Medical libraries, bioinformatics, and networked information: a coming convergence?

By Clifford Lynch, Ph.D.
Executive Director

Coalition for Networked Information
21 Dupont Circle, Suite #800
Washington, DC 20036-1109

Libraries will be changed by technological and social developments that are fueled by information technology, bioinformatics, and networked information. Libraries in highly focused settings such as the health sciences are at a pivotal point in their development as the synthesis of historically diverse and independent information sources transforms health care institutions. Boundaries are breaking down between published literature and research data, between research databases and clinical patient data, and between consumer health information and professional literature. This paper focuses on the dynamics that are occurring with networked information sources and the roles that libraries will need to play in the world of medical informatics in the early twenty-first century.

INTRODUCTION

This paper offers some brief thoughts on how the medical library in the twenty-first century may respond to the technological and social developments that are fueled by information technology, bioinformatics, and networked information. These developments will change both the library and its context. The practice, economics, and politics of biomedical research and health care delivery are already changing, and a significant part of this change appears in how patients, health care providers, researchers, policymakers, and the general public all relate to the corpus of biomedical information.

I believe that libraries that are part of highly focused settings (law, biomedical, and many corporate libraries), as opposed to those that serve a diverse community (research, academic, and public libraries), are at a pivotal point that will shape their future. The organizations in which they function are becoming profoundly information intensive. These libraries must decide whether or not to continue their relatively narrow, traditional role, in which they act simply as mediators of access to a limited body of published literature that comprises an ever decreasing proportion of the overall information environment. Their alternative is to take on a much larger and more dynamic role within their host organizations. These libraries may become extensively engaged with the processes and activities of the parent organization and the acquisition, capture, and management of information created by or used by that organization, even when such information falls far outside the published literature that has been the primary historical focus of most libraries.

As biomedical research and health care practice have come to rely on various types of databases (most notably in molecular biology and genetics, but also in clinical trials and public health), medical libraries have come to face such a decision, perhaps sooner than other types of special libraries. Not surprisingly, the biomedical library community served as the source and incubator for the concept of "knowledge management" [1]. Note that this concept should not be confused with the "knowledge management" that has become fashionable among management consultants in the late 1990s. In the late 1980s and early 1990s, this concept offered a vision of library engagement in broad research information management processes within host organizations. In hindsight, these ideas clearly should be extended to encompass operational information management, particularly given the increasingly competitive and commercialized nature of the health care system, the changing role of the patient or consumer, and the potentially more data-intensive modes of health care delivery that have been emerging.

I have tried to look beyond simple extrapolation of the historic role of biomedical libraries. I have tried to speculate broadly about the character and activities of libraries that have taken the second path and have become more active participants in biomedical research and health care delivery processes within their host

organizations. This paper will examine a series of developments involving different kinds of information. And it will discuss the sources, characteristics, use, and management of each, beginning with information that has traditionally been the clear province of the library, and moving outward into other classes of information that drive the processes of research, health policy, and delivery of medical care.

Let me emphasize that my training and experience is in computer science and library automation. I am not a physician, a biologist, or a librarian, and my experience in operational biomedical libraries is negligible. The reader should consider my comments as the speculations of a technologist.

PROFESSIONAL AND SCHOLARLY "PUBLISHED LITERATURE"

Clearly, both health care and research facilities will grow increasingly networked, both internally and in terms of connectivity to the global public Internet. As consolidations among health care providers and alliances and collaborations among researchers (including cross-sector alliances linking academia and industry) increase, electronic delivery of scholarly and professional literature will become increasingly necessary to respond both to the geographic scattering of participants and to the need for timely access to information.

The journals of the biomedical and health sciences are now mostly available in digital as well as print formats, though more as translations of the printed forms rather than any fundamental reconceptualization of either the contents or social processes of the traditional print scholarly journal. And they are often read on paper generated through distributed print-on-demand models. Interlibrary loan services are being largely replaced by the almost instant acquisition of articles on demand. Other materials (handbooks, textbooks in areas such as anatomy, and monographs) are being supplemented by electronic works that have been more fundamentally reconceptualized for the digital environment. Acquisition of the published literature by libraries has transformed into the negotiation of site licenses for network access to material.

The life and health sciences have historically been well served by abstracting and indexing resources, such as MEDLINE and BIOSIS Previews, which organize the disciplinary literature. These resources are now widely available online through interfaces suitable to end users rather than to trained searchers, and are becoming linked to the electronic forms of the journal literature. The journal literature itself is also becoming interlinked, permitting readers to navigate easily from citations to cited articles.

With the exception of nonjournal materials, which represent the beginnings of genuinely new genres of scholarly and professional communication, all of these

developments are direct translations of the existing print literature system into the digital domain. They exploit the independence of geography, near-instantaneous dissemination, and navigational flexibility that are intrinsic characteristics of this environment. They are not particularly surprising or revolutionary, although they will create some unexpected demands on the library, particularly in the negotiation and management of potentially expensive and complex licensing agreements, and the development and deployment of infrastructure in areas such as authentication and access management needed to support electronic access.

A cultural revolution is taking place in scholarly publishing. It is most obvious in physics and computer science (consider the Los Alamos physics preprint archive and the computer science technical report system, DIENST/NCSTRL) [2]. This shift values rapid and very democratic distribution of information outside the traditional channels of the peer-reviewed scholarly journal, by making it available to all interested parties on the Internet prior to peer review. The culture of preprints has never been strongly established in the biomedical and health sciences. In fact, editors of some biomedical journals are strongly opposed to prior distribution of results on the Internet. They favor the continuation of print practices, such as embargoes, arguing that the nature of the research forbids its release in unreviewed form or in a disorderly fashion. There are also patent and intellectual property issues that are more significant in the life sciences than in mathematics or physics, which argue for a very controlled and formal publication process. Such arguments reinforce the hegemony of the existing dominant journals. The recent series of *Science* articles [3] on the role of the embargo in scholarly publishing is an excellent example of such thinking. But to some outsiders, these arguments seem self-serving and biomedical journals appear to be simply self-satisfied, unrepentant reactionaries in a digital environment where many of the social practices of the scholarly journal are coming under intense questioning.

While biomedical publishing has been conservative, to say the least, change is clearly in the air. Harold Varmus, M.D., at the National Institutes of Health (NIH) has recently proposed a very large-scale Internet archive of biological and health sciences journal articles, and if this proposal moves forward, it promises to reshape scholarly communication in these disciplines [4].

This issue does not directly apply to biomedical libraries. They stand in a relationship to the literature as both funder and victim; but they have very modest influence in shaping the culture of scholarship in the life sciences. It is important to note that if a culture of rapid, democratic information distribution through preprints takes hold, the consequences for biomedical

libraries will be substantial. It will reduce the significance of the current canon of journals, and increase the importance of identifying and obtaining information moving through new distribution channels in a timely fashion.

CONSUMER HEALTH INFORMATION

The combination of structural and economic changes in the health care delivery system positions patients and consumers as adversaries to health care providers and insurance companies. As a result, there is a growing demand for consumer health information. This information permits patients to take greater charge of their own treatments, allows consumers to make better informed choices about their lifestyles, and helps families to deal with health matters. If a distrust of managed care, insurance companies, and medical establishment provides the motivation, the technologies of networked information provide the enabling tools. The public not only has access to information designed for the public, but access to most of the same information resources that are used by researchers and health care professionals. And the tools are not just limited to information access. They also incorporate new channels of interactive communication (electronic mailing lists, newsgroups, and bulletin boards), all of which can be used to share information, rumor, and hype among patients, consumers, and families, together with care providers, researchers, and assorted charlatans.

The network is a fantastically powerful vehicle for public access to information in personal finance and investing, areas that have historically been the domain of professionals. Consider the impact of the EDGAR database of Securities and Exchange Commission (SEC) filings, or the commoditization of near real time stock quotes—but consider also the impact of uncontrolled Web sites, mailing lists, and chat rooms offering what is sometimes very questionable financial and investing information. People expect that such open access and discourse should be available to health care information; and federal institutions such as NIH and the National Library of Medicine (NLM), as well as a range of universities, hospitals, commercial information providers, and quacks on the network are mobilizing to meet these consumer demands and expectations. As a result, heavily reviewed and carefully prepared offerings coexist with both regulated commercial presentations from the pharmaceutical industry and outrageous alternative perspectives from individuals and small businesses. Any content provider has an immediate global reach. The local hospital competes with national institutions and hospitals throughout the world for the attention of patients. Libraries in the future may provide evaluation and filtering of the torrent of consumer health information material avail-

able to the consumer, rather than simply increasing the volume of the torrent.

Historically, many biomedical libraries had little to do with communicating information directly to patients, consumers, and families. They provided professional and scholarly information to health care providers and researchers, and the health care providers were responsible for telling the patients what they thought the patients needed to know. To the extent that patients or the public wanted to consult the literature, the public library was left to mediate this process, though it held little of the relevant material from the scholarly literature. Now the situation is becoming more complex. Libraries are being asked to provide professional and scholarly information to the general public. To some extent, libraries have in fact been bypassed by the public availability—either for free or on a commercial basis—of references sources on the Internet that the public can consult without mediation. Patients are collecting abstracts from MEDLINE and demanding explanations from their physicians. And libraries are being asked to assist the public in understanding this material, which may be used as consumer health information if it is the only information available to the consumer. There are some very delicate practical and ethical dilemmas here: Should libraries facilitate access to professional literature that most consumers are not equipped to understand and use effectively, or ought they steer consumers away from this problematic but accurate professional literature towards sources designed for the lay public? Biomedical libraries will need to define the scope of their role as providers of public information and decide how much to invest in this kind of content.

There is another dimension to the consumer health issue. A number of institutions are beginning to provide high-quality consumer health information on the Internet. In effect, they have become publishers. This role is partially a public service; it is partially an extension of their health care provision to their local community; and it is, to some extent, advertising. This service could even be designed as a separate profit-making enterprise, much as some hospitals and universities publish health newsletters today. Whatever the motive, it casts the institution in the role of author and publisher. The issue here is the extent to which a biomedical library should involve itself in the content provision activities of its host institution, and how this service interacts with traditional library activities such as reference and collection development.

THE GRAY LITERATURE

Gray literature is normally defined as everything outside of the canon of published scholarly and professional literature (serving practitioners and researchers) and the popular press (serving consumers and, sec-

ondarily, the researchers and practitioners). Uniquely digital genres will be considered later in this paper. I will limit the present discussion to materials that have historically been distributed in print (although not necessarily formally published), including those that now may be distributed in analogs to their traditional print formats via the network (for example, press releases). This material has traditionally been of little significance to most biomedical libraries because their constituency has been care providers and researchers.

For most practitioners and researchers in health care delivery and biomedicine, the gray literature is of minimal importance unless there is a major cultural change in favor of preprints and other distributions of information outside of the traditional publishing channels. There are exceptions: announcements of clinical trials, literature from pharmaceutical companies, and guidance from insurance companies and managed care providers that may be relevant to physicians. For the practicing doctor, most of this material does not come through the library, I suspect, but arrives directly in a somewhat haphazard fashion. But to the extent that medicine is detective work (for example, in areas such as epidemiology), the gray literature is of significance. Perhaps it is most important to areas such as public health and health policy, where policy analyses; press releases; geospatial, environmental, and regulatory reports; laws; federal, state, and municipal regulations; and insurance provider policies are an important part of the knowledgebase.* Consumer health information and communication are also important components of the gray literature, providing a window into the practices and beliefs of the public. This information may have important implications for disease diagnosis and management strategies, or it may highlight potential problems with drug or lifestyle interactions.

The question for libraries, of course, is the extent to which they will acquire, organize, and deliver material from the gray literature, particularly as it moves to electronic form and becomes more difficult for the library's clientele to identify and use independently.

DATASETS, DATABASES, AND KNOWLEDGEBASES

The research literature today is complemented by datasets that provide the source evidence. In many cases, these datasets have to be deposited into databases or other repositories as a condition of publication so that they are available to other researchers. This condition is more and more the case in areas such as clinical trials, molecular biology, and X-ray crystallography of proteins.

These databases and repositories, however, have taken on lives of their own as community-wide consensus representations of the current state of research knowledge and its frontiers. They are, in effect, a second literature for the discipline, linked in increasingly complex ways to traditional journal literature and to abstracting and indexing databases that structure and organize that literature. Consider the work of the National Center for Biotechnology Information (NCBI) within NLM, as an example of how these linkages are developing. Also, these databases are updated continually, as activities such as genomic-sequencing work progress, as a direct reflection of research progress, rather than published merely as adjuncts to journal article publication.

In addition to such public, community-wide databases, the biotechnology industry produces innumerable specialized proprietary databases. These databases are usually licensed through collaborative agreements between the producers and other commercial entities, whether they are other biotechnology firms or large pharmaceutical corporations. But there are situations where these databases are also made available to the academic community, either under special research arrangements or as part of explicit collaborations. Collaborations may involve intellectual property licensing groups and legal counsel, as well as the researchers within an organization. And the role of the library here is an open question.

Databases are becoming increasingly important to practitioners as well as to researchers. There are databases to support the work of emergency rooms and poison control centers, and to codify best practices in the treatment and management of diseases. Databases listing current clinical trials that are underway and enrolling participants have become important tools for both primary health care physicians and patients. Drug interaction and pharmacy databases are an important component in managing care, particularly when multiple specialists are working with a patient.

It is important to understand the growing significance of these databases to various parts of the biomedical community, as well as their broader implications. Simply converting the journal literature into databases results in unstructured databases that rely heavily on language. Computers can do very little with them, other than to search them and retrieve results. Technologically sophisticated databases are developing that represent information in structured representations rather than simply as text. Computer programs can process and analyze the structured information, and then act upon the results of such computations— whether by making links, alerting humans, or refining new composite information. Researchers already operate in an ecology that includes computational agents as well as human analysts. Practitioners will increasingly move into this environment, too, as the infor-

* See the December 1998 issue of *Journal of Urban Health* for a series of papers on this topic.

mation and knowledgebases grow more structured and codified.

But how will the programs be managed and by whom? In research, these computational agents may simply make a connection between sequences or molecular structures that have been submitted to a database and alert the interested researchers. In health care delivery, the result may be identification of a relevant clinical trial starting up, of the availability of a new investigational drug, or of a change in the caregiver community assessment of the efficacy of a popular therapy. As these connections progress, the results may become even more complex: An evaluation made by an institutional computational agent, based on the integration of data from multiple sources, including proprietary outcomes data from within a health maintenance organization (HMO), might indicate that an investigational drug is more promising that previously believed. The development, deployment, and support of computational agents will be a critical information management function in the twenty-first century.

To the extent that these databases and knowledgebases are proprietary rather than public, medical libraries will be involved in negotiating and managing licensing agreements. Just as with the electronic literature, licensing agreements will need to be supported by an access management and authentication infrastructure. What role libraries will have in mediating access to these databases is less clear, given the advances in end-user-oriented interfaces. As researchers and even care providers develop programs that extract and correlate data from many sources, it will be important that they know what databases are available to be used, and under what legal and economic terms. Another question is the role that the library will have in teaching, particularly as these databases become an integral part of the curriculum for training both practitioners and researchers.

PUBLIC AND PROPRIETARY INFORMATION

Many sectors of the biomedical establishment need to maintain private information resources to complement those that are publicly available. They do this for many reasons: A research group may maintain private databases pertaining to research in progress that is still incomplete and not ready for publication. An HMO or other corporate care provider may maintain extensive data on best practices, outcomes, performance, and patient-related statistics, and may keep most of this information proprietary in the name of competitive economic advantage. Corporations and health care facilities need to manage the results of in-progress clinical trials. A useful distinction can be seen between raw information that will ultimately become public (in some refined form) and information that represents

competitive organizational advantage and that, from the organizational point of view, should remain proprietary and private indefinitely (although it may serve as source material for summaries or analyses that see public release).

Formal, highly managed processes exist for moving information from the private to the public sphere. There are patent filings, publication in the scholarly and professional literature (including the possible required filing of datasets) with the accompanying peer reviews, new drug approval processes, and release of information to public health and policymaking groups within the framework of legal or regulatory reporting requirements or reporting to insurance companies. Specific proprietary data (for example, success rates for treating certain types of illness or for some types of surgery) may be released as, in essence, advertising.

Intervention by the library in the organization of information from the research groups is very much in the spirit of knowledge management. With their organizational and information management skills, it is their contribution to the processes of research and knowledge dissemination. Management of corporate or organizational data is really much like what has traditionally been viewed as records management.

Digital technologies are positioning us to produce a vast stream of information products that can be transmitted and stored. It is not clear which of these products will be public, which will be commercial, and which will be private. In fact, ownership rights of these new information objects are often extremely unclear. Most are byproducts of either recording activities that take place in the physical world (meetings, conferences, seminars and symposia, hospital rounds, and surgical procedures) or of capturing activities that occur in new digital spaces (telemedicine, network-based consultations, collaborative data analysis and visualization, and network-based meetings involving teleconferencing). In either case, these activities have the potential to produce digital "documents" that can be archived, indexed, annotated, reviewed, and redistributed. These new digital documents are positioned at the intersection of scholarly communication, records management, and actual conduct of research and health care; and the roles of all interested parties (including libraries) in their management is unclear. But there is little question that they will be important—perhaps more so and sooner in health care and biomedical research than in many other disciplines.

Finally, there is a desire to integrate these private information holdings with public information resources in a unidirectional fashion. A researcher "on the inside" should be able to start with the proprietary resources and follow links to the public corpus of knowledge, literature, and data. But the links should be unilateral; the public and other researchers "on the outside" should not be able to access the proprietary

databases. This kind of information structure involves the dynamic re-presentation of public information in order to connect it to internal information resources. It also raises the possibility of developing bilateral agreements to exchange access to organizational "private space," and the need to be able to federate autonomous private information spaces. Libraries may be involved both in the negotiation of these agreements and their implementation.

PATIENT INFORMATION

The ownership of patient information is a complex and controversial topic [5]. I will only say that it is now a peculiar amalgam of information that belongs to or is shared among the individual patient, various interested organizational parties, and, to some extent, the public (usually in a redacted or statistical form). For care delivery and research organizations, it is a form of private record, but is generally handled by specialized patient information systems. Libraries, as far as I know, generally have very little to do with patient records.

Patient records are becoming richer and more digital as details are gathered into a central digital record and include various types of sensor scans (X-ray, MRI, EKG, etc.), as well as the usual lab test results and medical history. This trend is likely to expand enormously in the next ten to fifteen years. There are already pacemakers that are capable of uploading digital telemetry on a periodic basis and experimental "smart toilet" systems that can perform chemical analysis on human wastes and upload the results of this analysis. What has historically been a record of a person's encounters with the health care system can now become a regularly updated set of sensor readings that are collected from the person's environment, particularly as the sensors become more routinely integrated with the network. There is an enormous information management challenge implicit in these developments. Perhaps the most difficult issue will be determining when readings from the flood of data are anomalous and need to be brought to the attention of humans.

The integration of patient records with the biomedical literature (in the broadest sense) is going to be of increasing importance. We can view a patient record as a standing current awareness query on the part of the attending physicians, patient, and interested researchers against the various "literatures" described above. The explosion of medical information is making it increasingly difficult for physicians to remain current, and access to the latest relevant information will be crucial. Clearly, libraries will arrange access to many of the information resources for these current awareness searches. What is less clear is the role of the library in structuring and managing the search process.

Today, the primary search criteria of interest from patient records are likely to be medical conditions, symptoms, and medications prescribed. But in the coming era, molecular biology is going to be much more important and more personalized through approaches such as pharmacogenetics. Technologies such as genechips extract patient-specific genetic sequence information that can be used to help predict susceptibility to various diseases and perhaps anticipate disease onset, to calculate whether certain medications are likely to be helpful or harmful, and to predict susceptibility to environmental factors that may trigger disease [6]. Correlating this information with community-wide molecular biology databases will be essential both to its exploitation and to its interpretation. Over time, it will be useful to correlate this information to other diverse information sources that may provide information on factors such as patient exposure to environmental pollutants [7]. Patient records will expand to include structured databases that describe the patient at levels far beyond what is routine today. These databases will need to be linked and federated with reference databases, matched against the literature, and potentially mined by research programs.

And the results of these processes will need to be communicated to the patient on an ongoing basis. This process transcends records management, patient records, scholarly communication, or traditional functions of libraries and moves toward a new generation of integrated biomedical information systems that will take a far more holistic view of the relevant information universe, structured from a patient-centric perspective.

CONCLUSIONS

As I try to envision medical informatics in the early twenty-first century, the overarching theme is the synthesis of historically diverse and independent information sources. Boundaries are breaking down: boundaries between the published literature and research data, between research databases and clinical patient data, and between consumer health information and professional literature. New "digital documents" need to be placed within the information continuum alongside traditional forms of both literature and records. Such a synthesis is clearly important for research, for health care delivery, for formulation of health policy and practice of public health, and for individuals in understanding and taking charge of their own health. This synthesis faces formidable challenges: It requires linking public and private information and resolving complex privacy issues. And it requires economic frameworks that permit the sharing of commercial and noncommercial information, and that balance corporate commercial advantage with the public good.

This synthesis, no matter how important, will not just happen. And who is responsible for translating the vision into reality is not at all clear. Biomedical libraries clearly have a major stake in the process. They have traditionally acquired and managed key information resources that need to become part of this synthesis, and they have mediated and supported access to other component resources. Fully supporting the synthesis means being willing to see the biomedical literature that they have historically managed merge, at some level, into a much richer knowledgebase, but one that goes far beyond the normal concerns of libraries and their literature. Similar issues apply to the organizations that have managed patient record information and genomic databases; all must struggle with a potential loss of some control and of distinct, independent identity.

The issue is not how the Internet will change biomedical libraries, or even health care and biomedical research. The Internet provides connectivity for information access and information sharing, and this connectivity can only increase and improve. The Internet provides a context within which a whole series of developments in medical informatics, information technology, electronic commerce, molecular biology, sensors, and related areas can rapidly play out for biomedical research, health care, and the general public.

The shape of the future for the biomedical library, I believe, depends on its understanding of these broad secular trends, the insight and expertise that it can bring to advancing them, and its participation in the construction and operation of these new, very broadly based information systems that absorb libraries, records management, and patient records into an integrated framework. It depends on the ability and willingness of the library to step outside of its traditional roles and comfortable niches and to focus on larger, systemic goals and visions.

REFERENCES

1. LUCIER RE, MATHESON NW, BUTTER KA, REYNOLDS RE. The knowledge workstation: an electronic environment for knowledge management. Bull Med Libr Assoc 1988 Jul;76(3): 248–55.
2. LAGOZE C, DAVIS JR. DIENST—an architecture for distributed document libraries. Commun ACM 1995 Apr;38(4): 47.
3. MARSHALL E. Embargoes: good, bad or necessary evil? Science 1998 Oct 30;282:860–7.
4. MARSHALL E. Scientific publishing—Varmus circulates proposal for NIH-backed online venture. Science 1999 Apr 30;284(5415):718.
5. COMMITTEE ON MAINTAINING PRIVACY AND SECURITY IN HEALTH CARE APPLICATIONS OF THE NATIONAL INFORMATION INFRASTRUCTURE, COMPUTER SCIENCE AND TELECOMMUNICATIONS BOARD, COMMISSION ON PHYSICAL SCIENCES, MATHEMATICS, AND APPLICATIONS, NATIONAL RESEARCH COUNCIL. For the record: protecting electronic health information. Washington DC: National Academy Press, 1997.
6. SERVICE RF. Coming soon: the pocket DNA sequencer. Science 1998 Oct 16;282(5388):399–401.
7. YAP W, REJESKI D. Environmental policy in an age of genetics. Issues Sci and Technol Policy 1998 Fall;15(1):33–6.